Человек-вектор и подъем конверсии: как мы сделали шаг к Federated Learning

Анастасия Семенова
Иван Снегирев
Артём Просветов









О нас





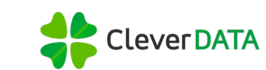
Анастасия Семенова

- Ведущий Data Scientist (CleverDATA)
- Веду семинары в ВШЭ

Артём Просветов

- Chief Data Scientist (LANIT/CleverDATA)
- Ведущий математик ИКИ РАН
- Веду лекции в ВШЭ
- к. ф.-м. н.





CLEVERDATA –
ТЕХНОЛОГИЧЕСКИЙ
ПРОВАЙДЕР РЕШЕНИЙ ДЛЯ
МОНЕТИЗАЦИИ ДАННЫХ



Data-marketing partners

dentsu AEGIS network

Distributors





Technology partners





Hermann.AI
by CleverDATA

AI



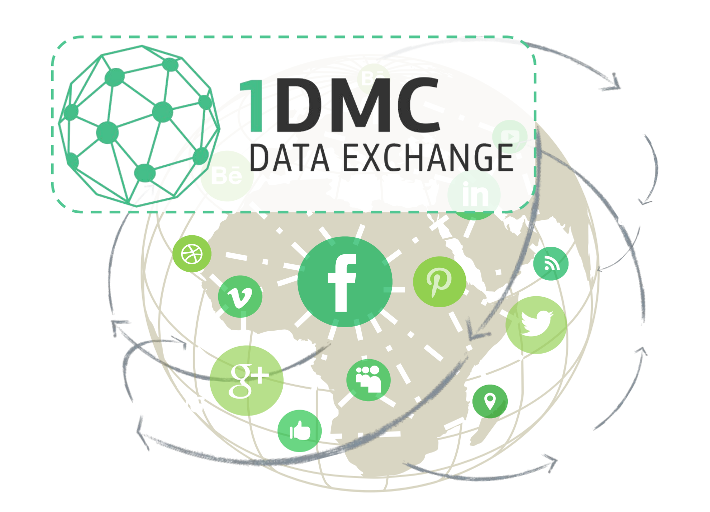
DATA EXCHANGE



BIG DATA

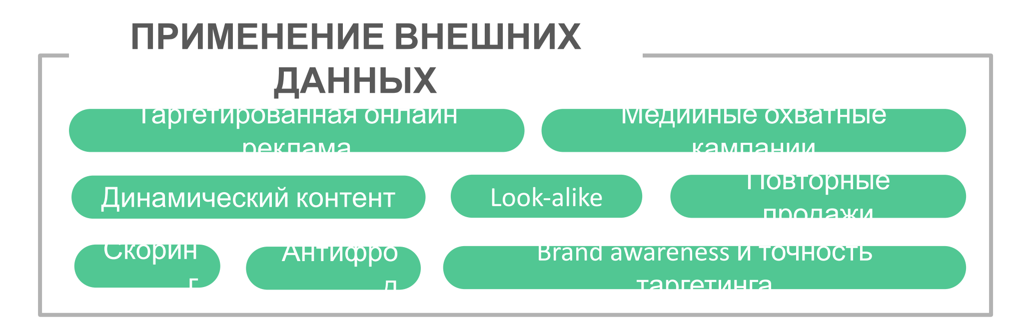


ЕДИНАЯ ТОЧКА ИНТЕГРАЦИИ ПОСТАВЩИКОВ И ПОТРЕБИТЕЛЕЙ ВНЕШНИХ ДАННЫХ



1DMC БИРЖА ДАННЫХ – площадка, объединяющая поставщиков и потребителей данных. Доступный объем данных сопоставим с крупнейшими в России сайтами.









Оглавление

- CleverDATA
- Задачи
- 2 проблемы
  - Текущее представление профиля человека и вопросы к сегментации
  - Что такое FL и почему он нам интересен
- Решение – эмбеддинги!
- Подход 1: Текст vs Граф
- Подход 2: Time Encoder





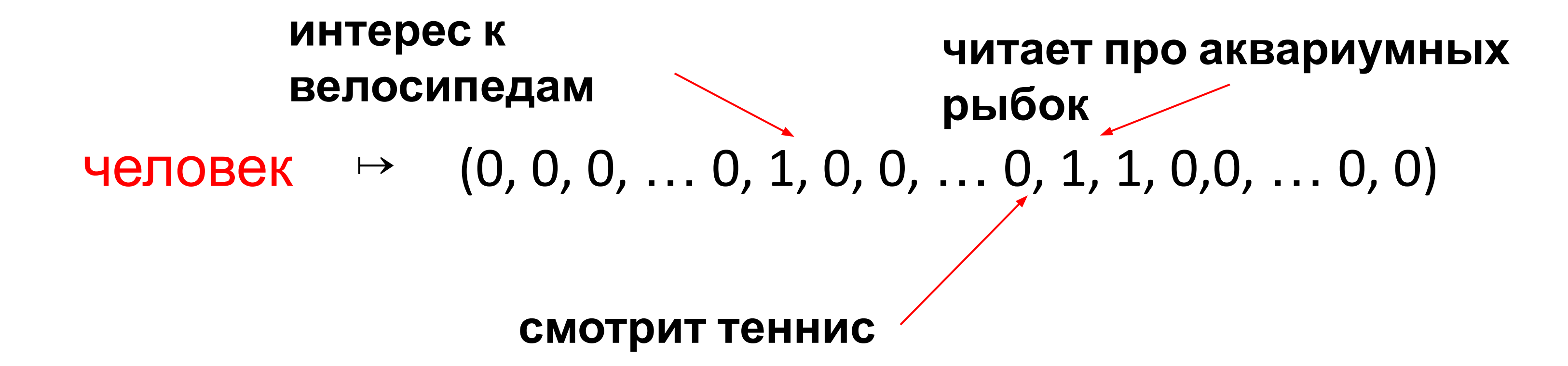
Задачи

1. Lookalike (LaL)
2. Вероятность целевого действия



Проблемы





HighLoad++ Весна 2021



Проблемы

1. Слишком длинный вектор профиля
2. Данные разрежены
3. Высокая размерность пространства представлений
4. Хотим Federated Learning





Federated Learning

Dataset компании 2

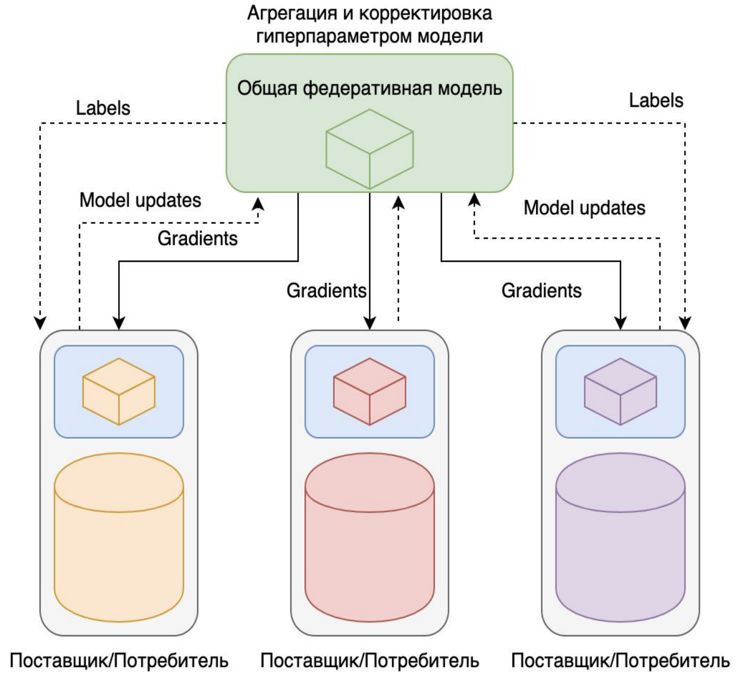
Labels компании 2

Labels компании 1

Dataset компании 1

Метки

Признаки



H Brendan McMahan Communication-efficient learning of deep networks from decentralized data.





1. Lookalike

Lookalike (LaL) – это таргетинг, при котором рекламные материалы показываются тем пользователям, которые по поведенческим характеристикам похожи на текущую аудиторию ресурса.

Clickstream – сегмент целевых пользователей (целевой класс)

Repr_sample – сегмент случайных пользователей

- tf-idf
- LogisticRegression / boosting / etc.





2. Предсказание осуществления целевого действия

Выбирается целевое действие на сайте клиента DMP.

Задача бинарной классификации: те, кто осуществил целевое действие (target) и кто – нет (not_target).





Проблемы

1. Слишком длинный вектор профиля
2. Данные разрежены
3. Высокая размерность пространства представлений
4. Хотим Federated Learning

Решение: Embeddings!





Пользователь1

Пользовательская история1

Page 1 View

Page 2 View

Page 3 View

Пользователь2

Пользовательская история2

Page 3 View

Page 1 View

Page 4 View

...

...

Page 1 View

Page X View

Page Y View

ПользовательN

Пользователская историяN

Page Z View

Page X View

Page 4 View

{[Пользователь1][0.01, 0.06, …, 0.07]}

{[Пользователь2][0.02, …,0.04]}

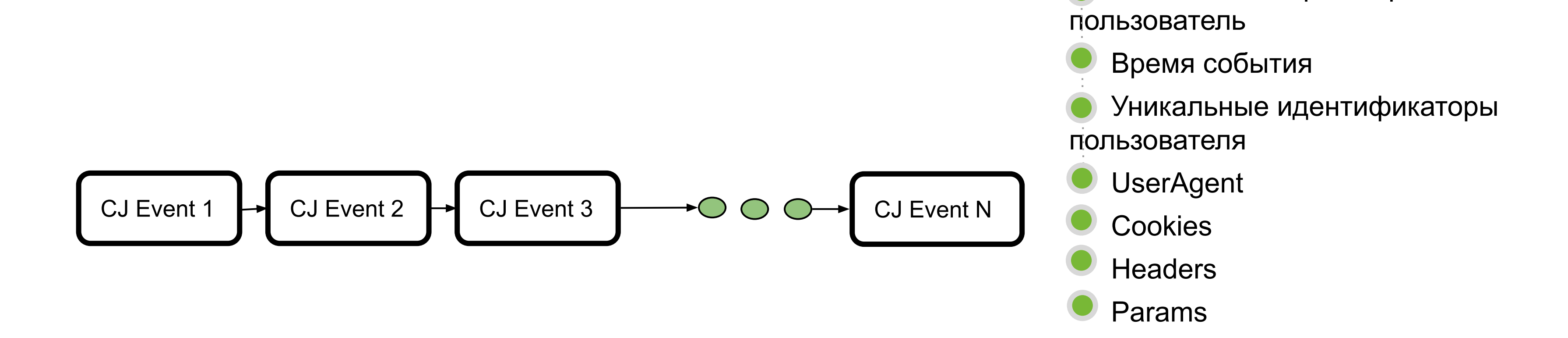
....

{[ПользовательN][0.08, 0.05, …, 0.97]}





Customer Journey (далее CJ)







Подготовка данных

user 1	URL 1	URL 2	URL 3	…	URL N
user 2	URL 1	URL 2	URL 3	…	URL N
…					
user N	URL 1	URL 2	URL 3	…	URL N





История первая

Текстовая векторизация



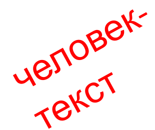


человек-
текст

Мир меняется… А ссылки?





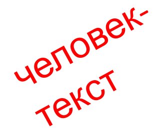


Мир меняется… А ссылки?

Был выбран FastText с буквенными n-граммами.







Мир меняется… А ссылки?

Был выбран FastText с буквенными n-граммами.

Пользователи = последовательности – это **документ**
Ссылка – это **токен**

user N	URL 3 , URL 1, URL 5, ⬭ ⬭ ⬭, URL N





человек-текст

INPUT PROJECTION OUTPUT

w(t-2)
w(t-1)
SUM
w(t)
w(t+1)
w(t+2)

or

INPUT PROJECTION OUTPUT

w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)

Given a set of (neighboring) words, **guess single words** that potentially occur along with this set of words.

CBOW

(Continuous Bag Of Words)

Skip-gram

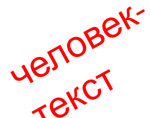
Guess potential neighboring words based on the single word being analyzed.

https://derekchia.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets/

https://mailsgun.ru/deep-learning-vs-common-sense-разрабатываем-чат-бота/







FastText от Gensim

Skip-Gram

dim = 128

min count token = 5

окно = 5





человек-
текст

user N

URL 3 , URL 1, URL 5, ⬤ ⬤ ⬤, URL N

model | model | model | ⬤ ⬤ ⬤ | model

вектор | вектор | вектор | ⬤ ⬤ ⬤ | вектор

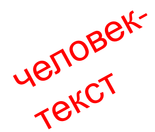
mean

user N

вектор

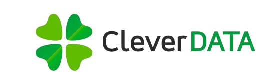


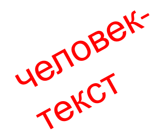




blabla_url.com *нет* в словаре





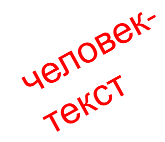


blabla_url.com *нет* в словаре

bla_url.com *есть* в словаре





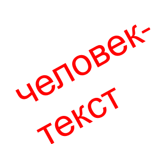


blabla_url.com *нет* в словаре

bla_url.com *есть* в словаре

n-gramm => :-) => есть вектор для **blabla_url.com**!





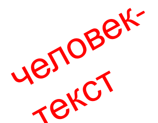


Плюсы

- ➔ Реализация Gensim позволяет обучаться на генераторах
- ➔ Метрики качества выше,чем у классического подхода
- ➔ Позволяет работать с незнакомыми токенами
- ➔ Большой прирост полноты относительно классического подхода



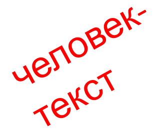




Подход	ROC AUC	F1-мера для целевого класса	Precision для целевого класса	Recall для целевого класса
classic	0.633	0.053	0.056	0.051
text_embs	0.664 (+4.9 %)	0.056 (+5.7 %)	0.032 (-42 %)	0.224 (+337 %)







Всё здорово, но чего-то не хватает

??? человек - не текст, человек - граф!





История вторая

Графовая векторизация



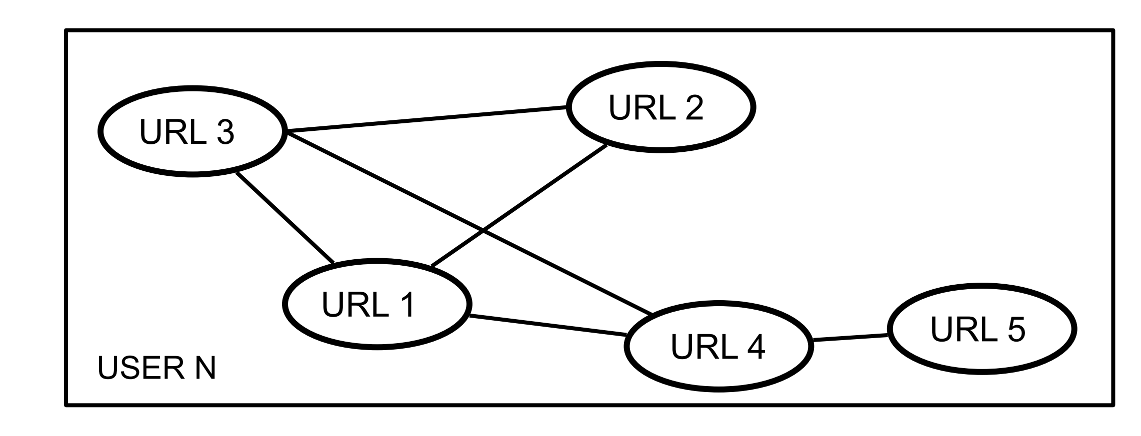




Структура переходов пользователей интернета важна!

Нижняя оценка размера графа на месячных данных составляет 160 000 доменных узлов.

При этом мы постоянно подключаем новых поставщиков данных.







человек-граф

Node2Vec

отображает соседние узлы в графе так, чтобы и в признаковом пространстве они были близки

Graph

sampling strategy

Input data

skip-gram model

Embeddings

Node2vec embedding process

https://arxiv.org/pdf/1607.00653.pdf
https://towardsdatascience.com/node2vec-embeddings-for-graph-data-32a866340fef





человек-граф

user N

URL 3 , URL 1, URL 5, , URL N

URL 3 — URL 1

URL 1 — URL 5

URL N-1 — URL N

model

вектор

model

вектор

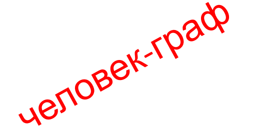
model

вектор

mean

user N

вектор

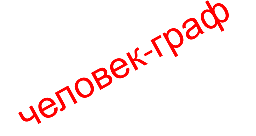


Плюсы



- ➔ Хорошая точность для уже известных токенов
- ➔ Время обучения не зависит от количества данных, только от размера графа
- ➔ Метрика полноты выше,чем у классического подхода



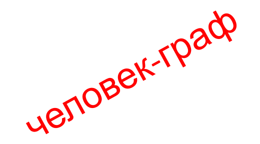


Минусы



- Нельзя рассчитать вектор для неизвестного токена
- Если у пользователя нет ни одного известного токена, то приходится брать среднее среди всех пользователей
- Рассчитываем вектор только для ребра, а не всей последовательности
- В целом показатели качества ниже, чем у классического подхода







Подход	ROC AUC	F1-мера для целевого класса	Precision для целевого класса	Recall для целевого класса
classic	0.633	0.053	0.056	0.051
graph_embs dim 128	0.627 (-0.9 %)	0.039 (-27.4 %)	0.023 (-58.7 %)	0.122 (+137.5 %)

?????? Метрики Precision и Recall получены в результате перебора порога бинаризации для максимизации F1-меры для целевого класса





человек – текст

или

человек – граф?







Теорема Кондорсе!





Теорема Кондорсе!

или по-простому: пробуем объединить подходы





текст (dim 128) + граф (dim 128) = emb (dim 256)





Подход	ROC AUC	F1 мера для целевого класса	Precision для целевого класса	Recall для целевого класса
classic	0.632 (0 %)	0.053 (0 %)	0.056 (0 %)	0.051 (0 %)
text_embs dim 128	0.664 (+4.9 %)	0.056 (+5.7 %)	0.032 (-42 %)	0.224 (+337 %)
graph_embs dim 128	0.627 (-0.9 %)	0.039 (-27.4 %)	0.023 (-58.7 %)	0.122 (+137.5 %)
embs_union dim 256	**0.679** (+7.3 %)	**0.062** (+16.6 %)	0.043 (-23 %)	0.115 (+125 %)

Метрики Precision и Recall получены в результате перебора порога бинаризации для максимизации F1 мера для целевого класса





Кейс

Онлайн-кинотеатр

Неглубокая полносвязная сеть

В 20 раз лучше, чем случайный предиктор





Лог -регрессия на сырых данных клистрим
Градиентный бустинг на сырых данных кликстрим
наилучшая модель на эмбедднгах

roc-auc

0,95
0,90
0,85
0,80
0,75
0,70
0,65

Сегмент посетителей онлайн-кинотеатра
Сегмент другого сайта





Что можно использовать еще?





Что можно использовать еще?

- Ссылка, с которой перешел пользователь
- Время события
- Уникальные идентификаторы пользователя
- UserAgent
- Cookies
- Headers
- Params





Что можно использовать еще?

- Ссылка, с которой перешел пользователь
- **Время события**
- Уникальные идентификаторы пользователя
- UserAgent
- Cookies
- Headers
- Params

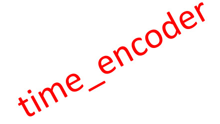




История третья

TimeEncoder

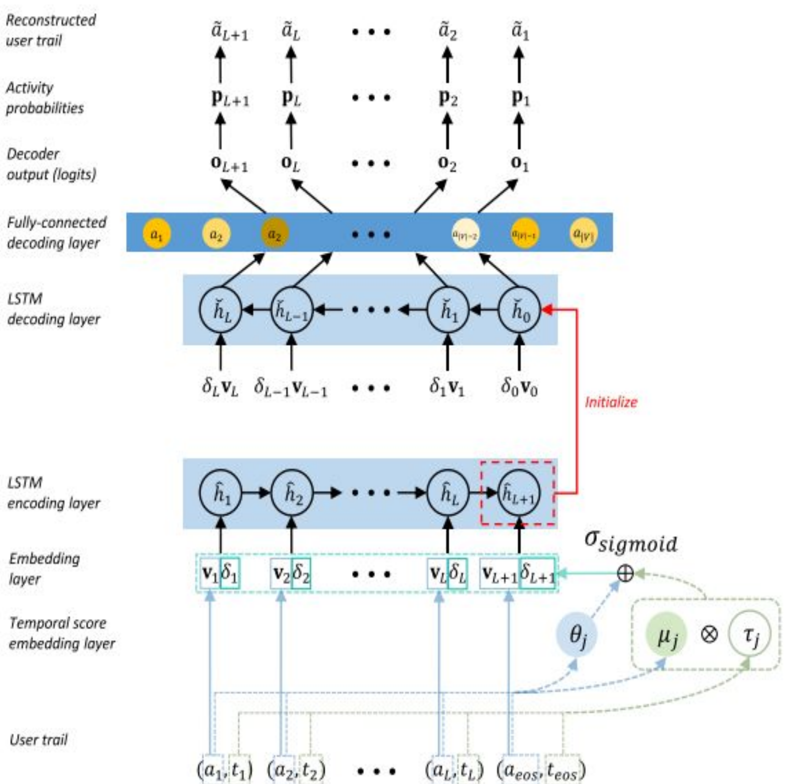






Sequence-to-sequence модель – это модель, принимающая на вход последовательность и возвращающая другую (такую же) последовательность элементов.











Параметры модели

- Размерность LSTMs, Embedding = 100
- Максимальная длина последовательности 64 (95% сессий имеют 61 и меньше событий в сессии)
- Количество эпох 10
- Оптимайзер Adam (lr = 0.001)
- Изменение CrossEntropyLoss в процессе обучения





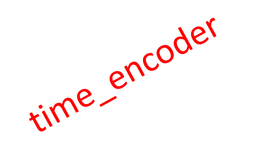
epoch
tag: loss/epoch

10.2
10.2
10.1
10.1
10.1
10.1
10.1
10.1
10.1
10.1
10.1
10.1
10.1

12 PM | 12 AM | 12 PM | 12 AM | 12 PM | 12 AM | 12 PM | 12 AM | 12 PM | 12 AM

March 2, 2021 | March 3, 2021 | March 4, 2021 | March 5, 2021 | March 6, 2021 | March 7, 2021



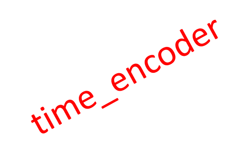




Векторизатор

- Для процесса получения векторов необходимо удалить декодирующую часть (DECODER) и оставить только кодирующую (ENCODER)
- Усредненный вектор выходов LSTM-слоя

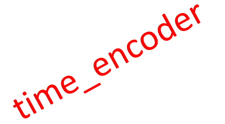






Как тестировали эмбеддинги?





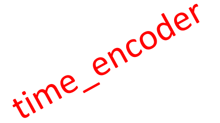


1 - Lookalike

Сравнение с нашим подходом, объединяющим **Текстовые** и **Графовые** эмбеддинги

Среди всех model ищется соответствующий метрике максимум, кроме случаев переобучения







1 - Lookalike

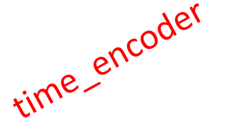
Сравнение с нашим подходом, объединяющим **Текстовые** и **Графовые** эмбеддинги

LaL CAR_X

Превосходство TimeEncoder над подходом graph_text:

- roc_auc **+22.5%**
- f1_score clickstream **+75%**
- f1_score repr_sample **+1.42%**







1 - Lookalike

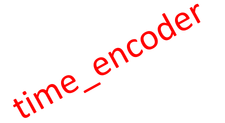
Сравнение с нашим подходом, объединяющим
Текстовые и **Графовые** эмбеддинги

LaL CAR_Y

Превосходство TimeEncoder над подходом graph_text:

- roc_auc **+8.23%**
- f1_score clickstream **+34.7%**
- f1_score repr_sample **+2%**







1 - Lookalike

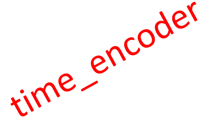
Сравнение с нашим подходом, объединяющим **Текстовые** и **Графовые** эмбеддинги

LaL APPARTS

Превосходство TimeEncoder над подходом graph_text:

- roc_auc **+6.5%**
- f1_score clickstream **+25%**
- f1_score repr_sample **+2%**









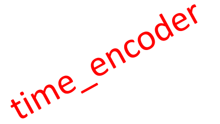
Сравнение с нашим подходом, объединяющим **Текстовые** и **Графовые** эмбеддинги

LaL AVIA

Превосходство TimeEncoder над подходом graph_text:

- roc_auc **+0.62%**
- f1_score clickstream **-3%**
- f1_score repr_sample **-0.12%**







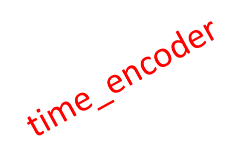
1 - Lookalike

Сравнение с нашим подходом, объединяющим **Текстовые** и **Графовые** эмбеддинги

Превосходство TimeEncoder над подходом graph_text (в среднем):

- roc_auc **+9.46%**
- f1_score clickstream **+32.93%**
- f1_score repr_sample **+1.325%**





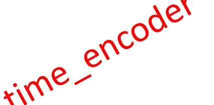


2 - Предсказание осуществления целевого действия

Превосходство TimeEncoder над подходом graph_text:

- roc_auc **+48%**
- f1-score target **+283%**
- f1-score not_target **+168%**
- f1-score_max for target **+281%**



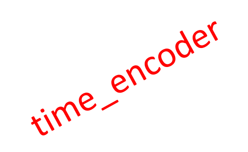




Дальнейшие шаги

- Изучить различные способы представления времени
- …
- А работает ли это на других данных?







Дальнейшие шаги

- Изучить различные способы представления времени
- …
- А работает ли это на других данных?

Clickstream

ОФД





Доставка - Стандартная доставка в течение 2-3 ...	0.00	2020-09-19 00:00:00
Бальзам для губ Lip Juicer Малина, свекла и им...	632.00	2020-09-19 00:00:00
Скраб для тела French Grape Seed	1592.00	2020-09-19 00:00:00
Йогурт для тела «Миндальное молочко»	792.00	2020-09-19 00:00:00
Пилинг для тела «Клубника»	712.00	2020-09-19 00:00:00





Данные ОФД

Текущий пайплайн работает хорошо

https://m.youtube.com/watch?v=R304Xy5eTso&index=105

https://habr.com/ru/company/lanit/blog/462959/





Данные ОФД

Текущий пайплайн работает хорошо

https://m.youtube.com/watch?v=R304Xy5eTso&index=105

https://habr.com/ru/company/lanit/blog/462959/

Хотим:

- Менее разреженные представления
- Неинтерпретируемые
- Вектор короче





Данные ОФД

Пилот с другим онлайн-телевидением:

Задача холодного старта рекомендательной системы





Данные ОФД

Пилот:

Проблема холодного старта рекомендательной системы

Успех! Ждём пресс-релиз







Эмбеддинги на данных ОФД

	Эмбеддинги
ROC-AUC	- 4%
Average Precision	+ 92%
NDCG	+ 40%
NDCG@50	+ 34%





Лёха,
где мы?

В пространстве
эмбеддингов

0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

−0.75 −0.50 −0.25 0.00 0.25 0.50 0.75





Выводы

Правильно подготовленные эмбеддинги способны на многое!

Подготовить наилучшим образом эмбеддинги помогут:

- статьи других исследователей
- добавление важной информации (например, времени)
- подходы из различных направлений
- эксперименты на различных задачах





Что-то новое может быть улучшенным старым

А может быть совсем новым :-)

Будьте открыты идеям и экспериментам!



https://t.me/Art_pr0



Спасибо!

Анастасия Семенова
Иван Снегирев
Артём Просветов





